DLA Piper LLP (US)
500 Eighth Street, NW
Washington, DC 20004
www.dlapiper.com

June 12, 2023

The Honorable Alan Davidson
Assistant Secretary of Commerce for Communications and Information
National Telecommunications and Information Administration (NTIA)
U.S. Department of Commerce
1401 Constitution Ave NW
Washington, DC 20230

Dear Assistant Secretary Davidson:

DLA Piper's global AI and Data Analytics Practice (DLAI) is pleased to respond to NTIA's AI Accountability Policy Request for Comment (Docket Number 230407-0093) to help inform policies related to the development of AI audits, assessments, certifications and other mechanisms to earn trust in AI systems. DLAI's response consists of 26 pages of targeted inputs on some of the most important outstanding questions related to AI accountability throughout the entire AI lifecycle.

Background on DLA Piper's AI and Data Analytics Practice:

Originally founded in 2019, DLAI is an industry-unique team of integrated lawyers, data scientists, and policymakers. Today, DLAI consists of over 100 lawyers, data scientists, coders, and policymakers focused on AI worldwide.

The group is chaired by Danny Tobey, who led Insider's 2022 list of top AI lawyers. Tobey sits on the Executive Committee of the UN's AI for Good legal track and is a founding member of the Health AI Partnership with the Mayo Clinic, Duke, and others. DLAI's Chief Data Scientist, Bennett Borden, joined the expanded practice in March 2023 and began his career at the Central Intelligence Agency (CIA), using data analytics and machine learning to describe, predict and influence human and organizational behaviors. Bennett was Financial Times's 2020 Innovative Lawyer of the Year, recognized for his "stand out expertise in data analytics and machine learning." Tony Samp leads AI policy for DLAI in Washington D.C. and was the founding director of the U.S. Senate Artificial Intelligence Caucus and a principal author of the Artificial Intelligence Initiative Act.

DLAI represents AI makers and adopters including the leading generative AI innovators, tests AI for trust, safety, and compliance, and builds AI-enabled tools to help respond to legal and compliance challenges. DLA's industry-leading AI practice also includes the team that patented Siri, the former Chief Privacy Officer of VISA and its Data Use Council chair, the former Chair of the Senate Intelligence Committee, the principal author of the first Federal Automated Vehicles Policy, and numerous other AI leaders. DLAI has drafted enterprise-wide Responsible

and Generative AI policies for some of the world's most recognizable brands and tested leading companies' AI for algorithmic bias and other compliance issues.

In 2021, DLAI participated in the NIST's Risk Management Framework process and built an "AI Scorebox" as a means to help organizations and businesses assess AI adoption readiness and governance maturity based on a series of questions and criteria. DLA Piper continues to stress that the involvement of legal, technical, financial, ethics, and senior leadership is critical for all organizations because the deployment of AI is something that needs to be assessed from many perspectives within an organization.

AI Accountability Objectives

1. What is the purpose of AI accountability mechanisms such as certifications, audits and assessments?

The main purposes of AI accountability mechanisms like certifications, audits, and assessments include:

- Verifying compliance with ethics, safety and fairness standards. Assessments help ensure AI systems meet established expectations around avoiding harm, bias and other potential downsides.

- Building public trust and confidence. Independent oversight demonstrates commitment to responsible AI practices and reassures stakeholders like customers, regulators and society as a whole.

- Identifying risks or issues proactively. Assessments can surface problems with data, model behavior or unintended consequences early before they become major incidents requiring intervention.

- Comparing to industry best practices. Certifications, audits and benchmarking against others developing AI systems promotes continuous improvement and adoption of protections that have become standard.

- Fulfilling legal and regulatory requirements. In some sectors like healthcare, finance and employment, laws may require validated AI safety, testing and review processes. External checks help meet those obligations.

- Managing organizational risk. Undertaking accountability mechanisms reduces potential liabilities in the event of accidents or AI failures by showing diligent governance and responsibility were exercised.

- Informing training and development. Insights from audits and assessments highlight areas to enhance AI through better data practices, debugging models and rethinking system design choices to improve.

- Providing transparency. Outcomes of accountability mechanisms offered publicly let people better understand an AI system's capabilities, limitations, and overall trustworthiness. Transparency builds confidence.

- Preventing misuse. Independent oversight serves as a deterrent to unethical or dangerous deployment of AI systems in ways that violate social norms and values. Accountability promotes responsible rather than reckless innovation.

AI accountability mechanisms aim to ensure safety and ethics, reveal any issues early, compare to industry best practices, help meet legal requirements, reduce organizational risks, fuel improvement, provide transparency, and prevent misuse. Overall, they serve to build public trust in AI's positive potential and mitigate any potential downsides through independent oversight and responsibility.

1.a. What kinds of topics should AI accountability mechanisms cover? How should they be scoped?

Some key topics that AI accountability mechanisms like audits and assessments could cover include are listed below. We note that not all of these topics are appropriate in all use cases or in all industries. The topics included in any specific use case should be tailored to the risks of that use case:

- Data practices - Reviewing the quality, diversity, and appropriateness of training data used for the AI system. Assessing potential issues like bias, inaccuracies, or underrepresentation of impacted groups that could skew system behavior.

- Model transparency - Evaluating the explainability of the system, including documentation of model features, logic, architectures, and development techniques. Assessing intelligibility and auditability.

- Testing rigor - Examining the thoroughness of testing methods used during development and continuously after deployment. Reviewing coverage of various use cases, edge cases, and potential failure or misuse modes. Checking that testing sufficiently simulates real-world conditions.

- Performance benchmarks - Measuring system accuracy, robustness, and reliability against trusted benchmarks and test data sets to verify competency at intended tasks. Comparing performance versus human baselines where applicable.

- Risk assessments - Reviewing process to identify potential risks and harms from the system, such as in areas of safety, fairness, accountability, privacy, security, etc. Evaluating the adequacy of safeguards implemented to mitigate those identified risks.

- Monitoring plans - Assessing ongoing monitoring for model errors, bias, deception attempts, abnormal usage patterns, and model drift from training. Reviewing triggers for investigation, escalation and response.

- Human oversight - Evaluating human involvement in system development, deployment and monitoring. Assessing appropriate human-AI collaboration, user controls, and expert judgment safeguarding critical functions or decisions versus full automation.

- Controls & safeguards - Reviewing architectural, technical and policy controls limiting the system's capabilities, access to data or systems, autonomy and potential harms. Assessing "fail safes" and overrides in case of issues.

- Handling failures - Reviewing processes for receiving complaints, investigating issues, suspending problematic uses, tracing root causes, notifying stakeholders, containing adverse impacts, and improving systems and controls to prevent recurrences.

- Transparency - Assessing system capabilities and limitations communicated to users and other stakeholders. Evaluating intelligibility of system actions and rationale to affected parties. Reviewing honesty in marketing and communication about its performance.

- Training for users - Reviewing education to users on appropriate, ethical ways to employ the system versus misuse. Assessing user awareness of risks, and training on using the system safely, effectively and responsibly.

Regular auditing and review across these areas of focus can verify AI systems meet key standards for responsible development and deployment. The specific topics and metrics assessed should align closely with the context and risks associated with the system's use cases as well. But covering these themes helps ensure safety, fairness, accountability and transparency.

1.b. What are assessments or internal audits most useful for? What are external assessments or audits most useful for?

Internal and external audits play complementary roles in providing oversight and accountability for AI systems.

Internal audits are useful for:

- Ongoing monitoring and quality assurance during development. Internal checks by the AI team can quickly identify issues to address iteratively before systems are deployed.

- Assessing compliance with internal policies and standards. Insights from internal reviews help ensure consistency with the organization's values and best practices for accountable AI.

- Building organizational culture around accountability. Regular internal reviews reinforce that oversight and responsibility are integral parts of day-to-day work.

- Developing baseline accountability before external review. Internal audits are a chance to resolve major issues and demonstrate due diligence before higher-stakes independent assessment.

- Informing training and process improvements. Internal insights direct focus to priority areas for improving team practices, tools, policies, and competencies to mature accountability mechanisms.

- Preparing for external audits. Internal dry runs of auditing build experience, test capabilities, and identify gaps to fill before official external reviews.

Independent, external audits are useful for:

- Unbiased assessment by specialists. External experts may identify problems overlooked by teams too close to the systems and provide an impartial perspective.

- Verifying compliance with laws and regulations. Independent validation helps prove alignment with legal and regulatory requirements around testing, bias, privacy, and other regulatory requirements.

- Reassuring customers and stakeholders. Trusted external validation helps demonstrate commitment to safety and responsibility to society. It promotes public confidence.

- Managing organizational risks. Independent oversight provides evidence of diligence in the event of any incidents and reduces liabilities associated with fielding AI systems.

- Comparing with industry standards. External audits benchmark organizations against peers, revealing how practices measure up versus established norms and best practices.

- Impartiality around sensitive findings. External parties may be best positioned to fairly judge and communicate on more serious issues uncovered without bias or conflict of interest.

- Spurring action on tough issues. The added weight of external audits prods organizations to address even difficult problems identified, rather than ignore them.

1.c. An audit or assessment may be used to verify a claim, verify compliance with legal standards, or assure compliance with non-binding trustworthy AI goals. Do these differences impact how audits or assessments are structured, credentialed, or communicated?

Yes, the specific purpose and nature of an AI audit or assessment can impact various factors in how it is structured, credentialed, and communicated:

- Verifying claims: Assessments focused on substantiating claims about an AI system's performance, capabilities or safeguards may involve more rigorous testing against benchmarks and standardized data sets, probabilistic validation methodologies, and head-to-head comparison with human experts. Results may be quantified and validated by statistics experts.

- Legal compliance: Audits checking compliance with regulations need to formally assess applicable legal requirements using accepted audit practices. Auditors require legal expertise or accreditation, or to partner with those who do. The process should closely track regulatory guidance with detailed pass/fail metrics and formal documentation suitable for regulatory review.

- Reporting: Audits related to legal requirements, claims or risks may require formal detailed reports with rigorous documentation of data, methods, scores/metrics, and any mitigation plans. Assessments against broader goals could produce higher-level reports for the public summarizing findings, recommendations and strengths/weaknesses in adherence to principles.

- Communication: Results with legal or liability implications would likely remain privileged, which could encourage companies to proactively investigate and remediate. Audit reports addressing organizational interests like public confidence or ethical standards may be publicly shared, or trends summarized without exposing confidential IP. Framing and transparency would differ by purpose.

Key distinctions in audit goals and standards - legal versus ethical, strict versus flexible, quantified versus qualitative, and so on - impact many factors in their structure, expertise required, documentation, reporting and communication style. Balancing rigor with practicality and transparency is essential to effective AI accountability.

1.d. Should AI audits or assessments be folded into other accountability mechanisms that focus on such goals as human rights, privacy protection, security, and diversity, equity, inclusion, and access?

There are good arguments for integrating AI audits and assessments into broader accountability mechanisms and review processes focused on goals like human rights, privacy, security, diversity and inclusion:

- Efficiency and reduced burden: Combining AI reviews into existing audit structures reduces duplication of efforts, minimizes disruption for organizations, and makes it easier to address multiple areas of risk or concern in a consolidated way.

- Holistic oversight: Assessing AI systems alongside other programs, policies and practices provides more complete, holistic oversight and ensures AI is not evaluated in isolation from larger organizational impacts on people. Interconnected risks and mitigations may be identified.

- Centralized expertise: Existing teams specializing in human rights, privacy, security, etc. can build AI expertise to handle integrated reviews, rather than organizations needing to hire separate dedicated AI auditors. Consolidated auditing leverages current resources.

- Clearer incentives: Including AI assessments in standard review processes pushes AI teams to prioritize broader values like inclusion, accessibility and transparency, not just technical benchmarks. It incentivizes developing accountable AI in balance with other organizational goals.

- Consistent standards: Integrated audits enable consistent criteria and expectations to be applied to AI systems and conventional programs. AI should uphold the same standards of privacy, ethics and accountability reflected across the organization's culture, not be treated as an exception.

- Shared lessons: Insights and improvements prompted by examining AI systems alongside other tools and processes promote shared learning. Best practices diffuse more readily when assessments are coordinated.

- Stakeholder alignment: A unified review structure reinforces that AI is not developing in isolation - it should serve all stakeholders just as other organizational programs aim to benefit diverse groups. Integrated audits reinforce this alignment.

- Signaling commitment: Including AI reviews in existing auditing demonstrates the organization's commitment to responsibility, safety and ethics in AI as integral to its overall standards, not a disconnected domain with separate rules. It signals the "right values" are being proactively operationalized.

That said, there are also arguments for independent AI-specific audits in some contexts given the unique nature, complexity, and risks associated with AI systems. But overall, combining AI assessments into existing accountability structures where possible has many advantages and should likely be the default model. AI should be held to the same high standards as other programs that serve people and communities.

1.e. Can AI accountability practices have meaningful impact in the absence of legal standards and enforceable risk thresholds? What is the role for courts, legislatures, and rulemaking bodies?

AI accountability practices like audits, disclosures and risk assessments can certainly have meaningful positive impact on their own through:

- Promoting responsible culture. Accountability mechanisms foster awareness, due diligence and ethical values even without legal standards enforcement. Cultural norms emerge around oversight and care.

- Building public understanding and trust. Voluntary transparency, open audit results and honest communication around limitations increase public confidence in AI despite lack of binding rules.

- Incentivizing continuous improvement. Accountability practices motivate those building AI to continuously enhance practices as issues are uncovered or norms evolve. Desire for strong reputation spurs progress.

- Allowing flexibility and speed. In fast-moving technical domains like AI, accountable practices adopted flexibly may outpace what regulation can mandate, allowing more agile development.

- Developing consensus solutions. Widely adopted practices could form the basis for sensible regulation or standards later as consensus on approaches emerges from experience.

However, meaningful risks likely still remain without enforceable legal standards around areas like safety, testing, algorithmic transparency, bias prevention and mitigating harmful impacts. Key roles for lawmakers and oversight bodies include:

- Setting liability rules. Courts shape precedent around accountability for harm and influence developer behavior through risk of liability suits or fines for issues like injuries, discrimination, violations of due process, etc.

- Enacting safety regulations. Legislatures directly regulate high-risk areas like medical diagnosis, autonomous vehicles, credit lending, etc., to mandate safety practices, validation, and transparency when harm could ensue.

- Establishing rights and protections. Enforceable guarantees around privacy, due process, non-discrimination, consumer protection and other areas uphold fundamental values that detailed practices should fulfill. Statutes, rules and regulations provide the framework.

- Subpoena power over opaque platforms. Regulators can compel platforms to share details on AI systems impacting the public interest in a way that voluntary disclosure norms may not reveal.

- Mandating disclosure in consumer services. Laws can require AI providers be transparent about use of automated systems, capabilities, limitations, etc., to address information imbalances.

- Facilitating civil society input and oversight. Formal mechanisms for public consultation, input and oversight of AI systems by impacted groups and advocates help ensure broader needs are met, not just those of developers and users.

- Tailored laws and regulations can be right-sized by risk level and use case to foster innovation.

AI accountability practices provide tremendous value - but likely achieve their deepest impact and protect against exploitation when backed by the force of law, regulatory oversight, public input, and enforcement by courts and lawmakers acting to defend rights and the public good. The two approaches are complementary.

2. Is the value of certifications, audits, and assessments mostly to promote trust for external stakeholders or is it to change internal processes? How might the answer influence policy design?

The value of AI accountability mechanisms like certifications, audits and assessments stems both from:

- Promoting trust and confidence externally in stakeholders like customers, regulators and the general public. And
- Influencing positive changes internally in organizational processes, culture and system design.

External value:

- Reassuring customers/users about safety, quality and ethics.
- Verifying claims for stakeholders like business partners.
- Demonstrating compliance to avoid penalties from regulators.
- Protecting reputation and brand image publicly.
- Deterring misuse by employees by signaling diligence.

Internal value:

- Incentivizing investments in accountability throughout design.
- Fueling cultural norms of responsibility and ethics.
- Enabling continuous improvement as issues are found.
- Promoting adoption of best practices and controls.
- Encouraging transparency and openness internally.
- Training personnel on oversight processes.

So, to influence policy design, governments and standards bodies should consider:

- Balancing both external and internal value. Accountability policies that benefit one but not the other may fall short.
- Using voluntary incentives before mandates if possible. Incentives allow flexibility for internal changes to emerge alongside external signaling.
- Mandating accountabilities cautiously. Strict top-down requirements could discourage innovation without the right internal commitments already in place.
- Promoting cultural norms first. Laws and rules may work best when reinforcing shared values that grow organically, not replacing them. Soft power alongside hard power is ideal.
- Allowing accountability mechanisms to inform regulation. Insights from voluntary practices could shape pragmatic and sensible formal requirements rooted in experience.
- Fostering both symbolic and substantive change. The ideal policy solutions couple strong external signaling around responsibility with deep internal impacts on culture and practice.

In essence, a healthy policy ecosystem likely balances mandatory accountability mechanisms where risks demand it with voluntary incentives and platforms to share best practices. The goal is external trust and verifiability combined with internal ethical norms and willingness to change. Striking that balance helps accountability mechanisms achieve their fullest potential impact.

3. AI accountability measures have been proposed in connection with many different goals, including those listed below. To what extent are there tradeoffs among these goals? To what extent can these inquiries be conducted by a single team or instrument?

There are certainly some inherent tensions and tradeoffs that could arise, but also opportunities for synergy:

Potential tradeoffs:

- Privacy vs transparency - More transparency about system logic/data may improve contestability but infringe on privacy and intellectual property rights.

- Non-discrimination vs safety/accuracy - Removing attributes that could enable bias could reduce predictive accuracy for sensitive use cases
- Explanation vs effectiveness - Increasing intelligibility for users often requires simplifications that reduce overall performance
- Human oversight vs automation benefits - Preserving human discretion limits speed and scalability advantages
- Transparency vs. security - The more transparent a model is the more susceptible it is to bad actor manipulation.
- Sensitivity vs specificity – Society may demand a different balance or false positives and false negatives depending on context and use case.

Potential alignments:

- Privacy, security and safety controls often reinforce each other
- Mitigating risks requires transparency for human oversight
- Avoiding discrimination/bias aligns with improving safety and legal compliance
- Consultation mechanisms aid redress and consent processes
- Shared technical infrastructure can enable various forms of auditing

For a single team conducting reviews, tradeoffs may be manageable through:

- Expertise in key disciplines like business expertise, law, ethics, technology, social science
- Engineering tools to balance competing goals where possible
- Consulting stakeholders and affected groups
- Contextual analysis of specific use cases and populations served to understand the risks and benefits of deployment

But oversight across the full lifecycle and set of concerns likely requires coordination between stakeholders with different specialties. A collaborative, multidisciplinary approach to

accountability balances competing goals and identifies positive synergies. Technical, ethical, legal, business and user perspectives should all inform good governance.

4. ***Can AI accountability mechanisms effectively deal with systemic and/or collective risks of harm, for example, with respect to worker and workplace health and safety, the health and safety of marginalized communities, the democratic process, human autonomy, or emergent risks?***

Yes. The overarching purpose of AI accountability mechanisms is for companies deploying AI systems to identify the risks they are introducing into the market or, for internal use, into its workforce, and establish mechanisms to reasonably mitigate those risks. DLAI has helped companies across the globe and in multiple industries effectively identify and mitigate the risks of AI systems, including systemic and collective risks.

For example, AI in the healthcare sector deserves special attention. Healthcare bias can be different from other bias. Unlike in other industry sectors where it may be advantageous to eliminate traditional categories of bias (particularly for protected classes like race and gender), in the context of healthcare, "bias" along those lines could be potentially relevant for detection and evaluation of medical conditions which can impact patient care and outcomes. However, if left unchecked, AI in digital health can scale and amplify historic and existing discrimination and disparities that are unfounded and irrelevant to patient care in a potentially unique, and harmful, way. In fact, there are reports of how bias in medical AI has resulted in exclusion of certain populations from care management programs or how algorithms designed to detect diseases such as cancer that were not trained on wide enough data sets representing a diverse population, performed worse for minority populations.

The critical consideration for assessing and auditing AI systems in healthcare is the need to understand how data inputs will be used and applied to prevent impermissible or unintended bias. Such assessments and audits need to be conducted at several stages to best ensure that algorithms are designed, trained, and applied in ways that do not perpetuate and amplify pre-existing bias. A foundational step to preventing and mitigating bias in digital healthcare is first understanding the potential paths for bias to get into AI systems in the first place. For example, bias can be introduced at the initial input stage where missing or underrepresented data sets or the inclusion of unintended signals based on proxy variables can impact how an otherwise bias-free algorithm functions. Bias can also be built into the model design or training process itself. Finally, models may drift over time, so even models that initially function without bias can shift without proper vigilance. Accordingly, deployment of effective audits for AI systems used in

healthcare algorithms should be mindful of these potential sources of bias and take affirmative steps to help mitigate and eliminate impermissible discrimination.

In addition, audits and assessments of AI deployed in a clinical setting, i.e., to treat, diagnose, or provide treatment recommendations, must be tailored to address the intersection of clinical and digital risks. The degree of clinical risk will be informed by the severity of the illness the AI is designed to address, as well as the extent to which a health care provider is relying upon AI to drive clinical decision-making. In other words, AI tools used to diagnose, treat, or address severe or critical illness generally pose higher clinical risks, and audits and assessments of those tools can be adapted accordingly. Explainability can be especially important in the healthcare setting, as the traditional healthcare model puts healthcare providers in the position of communicating risks and medical information to patients. As healthcare providers adopt AI tools into their practice, assessments of these tools should evaluate the degree to which providers can accurately and appropriately understand and communicate the information, recommendations, and/or diagnoses they generate.

Audits and assessments of AI used in clinical decision making should be performed under clinical conditions. Testing should evaluate not just the tool in isolation, but perhaps most importantly, how the healthcare provider – AI team performs collectively, in a real-world setting. Assessments should take into account traditional health system workflows and standard practices to ensure that the AI tool conforms to, and does not conflict with, established medical guidelines and practice standards.

5. ***Given the likely integration of generative AI tools such as large language models (e.g., ChatGPT) or other general-purpose AI or foundational models into downstream products, how can AI accountability mechanisms inform people about how such tools are operating and/or whether the tools comply with standards for trustworthy AI?***

Regulatory focus should be on where an AI system impacts a customer, employer, user, etc. These touch points are measurable and objective standards of accuracy, fairness, bias and harm can be applied to them. Often, an entity's touch points with a customer, employee or user, whether a product of AI or human decision making, is where any harm should be measured.

The disclosure of the use of AI need not be required in all cases. If the AI system is accurate, fair, and safe, with compliance monitoring in place and no risk of confusion or deception, then it may not matter that it is being used in a business process.

6. The application of accountability measures (whether voluntary or regulatory) is more straightforward for some trustworthy AI goals than for others. With respect to which trustworthy AI goals are there existing requirements or standards?

Areas where existing standards and precedents are more robust include:

- Safety and Legal Compliance - Regulations for product safety, medical efficacy, financial controls, etc., provide frameworks to audit against. Liability laws also shape accountability.

- Privacy - Data protection laws like state privacy laws, the FTC's rules on unfair and deceptive trade practices, and the GDPR provide concrete rules around data handling that enable audits, plus privacy-preserving techniques are advancing.

- Algorithmic Transparency - Techniques for model explanation and standards around documentation are maturing to improve interpretability and auditability.

Areas with fewer established standards and frameworks include:

- Avoiding Broader Harms - Systemic issues like effects on human autonomy, emergent risks, political impacts, workplace displacement etc. are hard to define, measure and regulate.

- Equity and Accessibility - Principles around universal access, inclusion and avoiding marginalization are important but can be difficult to operationalize and audit concretely.

- Democratic Oversight - Mechanisms for genuine public scrutiny, input and even veto over technologies are challenging to implement, though important to explore.

- Managing Tradeoffs - There are few set standards for balancing safety with privacy, transparency with effectiveness, and other goals, requiring context-specific analysis.

Safety, legal compliance, privacy, and transparency have more developed foundations for accountability than broader issues of systemic impacts, accessibility, democratic control and managing goal tradeoffs.

Though the latter areas lack concrete standards, that makes building consensus and governance guardrails around them all the more crucial. Accountability starts from norms and dialogue. But creative, collaborative policymaking is needed to realize AI's benefits while avoiding unintended

harm. Setting auditable requirements even where fuzzy is better than leaving rules undefined. Progress will come gradually through experience.

6. ***[continued]... Are there any trustworthy AI goals that are not amenable to requirements or standards? How should accountability policies, whether governmental or non-governmental, treat these differences?***

The core goals around developing trustworthy AI systems with positive societal impact should involve some element of requirements or standards to help ensure responsibility and safety. However, certain goals may be more challenging to define concrete accountability measures for:

- Avoiding broader harms - Assessing and regulating broad, systemic impacts on issues like human autonomy, emergent risks, political processes, economic displacement, etc., is extremely difficult due to the complex, unpredictable nature of such effects. Clear standards are elusive.

- Managing goal tradeoffs - Balancing competing objectives like transparency vs privacy, or safety vs effectiveness responsibly on a case-by-case basis depends heavily on context and involves subjective judgment. Universal standards are hard.

- Promoting societal benefit - Defining what constitutes truly "beneficial" AI, beyond simply avoiding potential downsides, is contentious. Reaching consensus on positive goals to encode into standards is less straightforward.

- Assessing future risks - Anticipating hypothetical longer-term risks from AI systems, like the existential threat of superintelligent AI, remains speculative. There is no clear foundation yet for concrete requirements.

So, while every trustworthy AI goal deserves attention, some may need to rely more on broader vision, values and incentives rather than defined standards. The most complex, systemic and long-term issues will require other solutions like democratic oversight, stakeholder engagement and ongoing debate on risks versus benefits.

But open, creative policymaking and public input can gradually build some consensus on reasonable precautions and review processes even for fuzzy goals. The conversation itself raises awareness and pushes ideas forward. There are always opportunities to develop "soft guardrails" if not always enforceable mandates.

Accountability policies around AI, whether from government or other entities, should aim to take a balanced approach that recognizes differences in maturity and feasibility across various goals and focus areas:

- Establish clear standards where possible: For goals with existing legal/regulatory precedents like safety, privacy, and non-discrimination, policies should codify auditable requirements appropriate to the context. Build on foundations in place.

- Promote further dialogue and research: For complex goals like managing tradeoffs, gauging risks, and defining societal benefits, policies should fund ongoing studies and convene experts from diverse disciplines to develop norms and potential oversight approaches.

- Create incentives for responsible practice: For organizational culture and behaviors, policies should encourage voluntary initiatives around ethics and accountability through recognition, investment, and flexibility to experiment. Let norms emerge alongside rules.

- Begin developing consensus "soft guardrails": For challenges like systemic harms or future risks, articulate principles and advisory guidelines that begin to outline reasonable precautions and oversight. Aim for coherence and direction.

- Contextualize policies for different applications: Account for unique risks, priorities and feasibilities between consumer AI versus high-stakes government or industrial uses. Calibrate policies to applications.

- Collaborate broadly across sectors: Include perspectives from technologists, ethicists, lawmakers, users, impacted groups and civil society to balance interests. Foster shared responsibility.

- Regularly review and refine policies: Adapt policies to the pace of AI progress, lessons learned, and new challenges that arise. Take an iterative, open and participatory approach to complex issues.

Overall, policies should combine clear mandates where possible with investments to advance the conversation and build consensus where current uncertainty remains. Patience, flexibility and principles-focused "soft guardrails" can lay the foundations for healthier norms and feasible standards to crystalize over time. But ambitious, creative policymaking efforts are needed now to keep pace with technology and steer it toward societally beneficial ends.

7. Are there ways in which accountability mechanisms are unlikely to further, and might even frustrate, the development of trustworthy AI? Are there accountability mechanisms that unduly impact AI innovation and the competitiveness of U.S. developers?

Accountability mechanisms do have the potential to constrain AI innovation if not designed thoughtfully. A few key risks include:

- Slowing pace of progress - Overly strict testing, validation and documentation requirements could significantly slow experimentation and the pace of beneficial advances.
- Stifling creativity - Highly prescriptive rules around techniques, architectures and use cases could inhibit new ideas and approaches from emerging organically.
- Impeding competitiveness - Strict requirements that highly disadvantage US developers versus peers internationally could hamper innovation leadership.
- Diverting resources - Significant costs in time, funding and talent to fulfill accountability demands reduces investment in advancing core capabilities.
- Fueling distrust - Focusing predominantly on hypothetical risks instead of actual benefits could undermine public confidence in and support for AI innovation.
- Ossifying standards - Inflexible, complex rules could fail to adapt to AI's quick progress and create outdated, irrelevant constraints on cutting-edge applications.

To mitigate these risks, policymakers, regulators and standards bodies should:

- Balance accountability rules with flexibility - Avoid highly rigid requirements that constrain experimentation, allow customization to context.
- Favor principle-based guidelines - Adopt "soft guardrails" allowing room for interpretation, rather than excessive technical prescription.
- Align international standards - Coordinate globally to minimize compliance gaps disadvantaging US developers.
- Consider costs holistically - Weigh obligations for developers against timelines and resources available. Account for tradeoffs.

- Enable collaboration - Ensure transparency demands allow latitude for sharing around benchmarks, methods and lessons learned to fuel collective progress.

- Address highest risks judiciously - Prioritize oversight proportional to realistic dangers posed versus precautionary limitations on theoretical risks that could undermine confidence.

- Emphasize ethical culture alongside rules - Recognize organizational values matter as much as formal compliance. Incentivize conscience and responsibility.

- Consider how policy can lift regulatory hurdles to facilitate the creation of reliable and representative datasets, allowing AI developers to access higher quality sources of data, improve AI outcomes, and reduce bias.

Ideally, accountability guardrails and incentives can accelerate progress by fostering trust but do run risks of frustrating innovation if not designed with care and wisdom. Ongoing dialogue across stakeholders is key to reaching policies that responsibly enable AI advances.

16. The lifecycle of any given AI system or component also presents distinct junctures for assessment, audit, and other measures. For example, in the case of bias, it has been shown that "[b]ias is prevalent in the assumptions about which data should be used, what AI models should be developed, where the AI system should be placed—or if AI is required at all." [82] ***How should AI accountability mechanisms consider the AI lifecycle? Responses could address the following:***

16.a. Should AI accountability mechanisms focus narrowly on the technical characteristics of a defined model and relevant data? Or should they feature other aspects of the socio- technical system, including the system in which the AI is embedded? When is the narrower scope better and when is the broader better? How can the scope and limitations of the accountability mechanism be effectively communicated to outside stakeholders?

AI accountability mechanisms should be focused on where risks are reasonably likely to be introduced by the specific use and context of the AI system. At times, those risks will be narrowly confined and at other times more systemic or contextual. Some examples of narrowly confined AI systems are:

- Highly specialized research systems intended for limited academic use rather than real-world deployment. Assessing societal risks may be less relevant than core functionality.

- Internal tools for tasks like data processing, analytics or simulation where human contexts of use and potential harms are minimal. Technical soundness may be the priority.

- Infrastructure components like computer vision APIs intended for incorporation into a wide range of unpredictable end applications. Broad impacts are hard to predict.

- Models shared on open repositories for general use rather than a specific product or service. Downstream usage patterns can't be fully known.

- Low-risk chatbots, digital assistants or recommendation engines meant for benign consumer domains. Few realistic avenues for serious harm exist.

- Proof-of-concept prototypes exploring new techniques where eventual applications remain uncertain. Hard to speculate on downstream risks.

- Systems under development for domains where extensive external regulation, testing and approval will be separately required prior to deployment. Redundant to assess highly speculative risks too early.

In general, specialized systems intended for controlled environments or academic progress, as well as building block components that enable a huge range of uses, may warrant a more focused risk lens. The same is true where realistic harm potential is low.

16.b. How should AI audits or assessments be timed? At what stage of design, development, and deployment should they take place to provide meaningful accountability?

Periodically and at critical milestones such as:

- Prior to initial model development to establish baselines

- During data collection, processing and labeling

- Following initial training and testing

- Before promotion to key decision-making roles

- After deployment to real-world conditions

- On a regular basis aligned to usage and risk levels (e.g., annual high-stakes reviews)
- Requiring prompt assessment after incidents, safety events or other triggers indicating potential issues

19. As governments at all levels increase their use of AI systems, what should the public expect in terms of audits and assessments of AI systems deployed as part of public programs? Should the accountability practices for AI systems deployed in the public sector differ from those used for private sector AI? How can government procurement practices help create a productive AI accountability ecosystem?

To the extent that public and private AI systems are used to make decisions about human subjects, the requirements should differ. The auditing requirements for an AI system should depend on the inherent risk to individuals, groups, and societies, not on who deploys the AI system.

Accountability Inputs and Transparency

20. What sorts of records (e.g., logs, versions, model selection, data selection) and other documentation should developers and deployers of AI systems keep in order to support AI accountability? How long should this documentation be retained? Are there design principles (including technical design) for AI systems that would foster accountability-by-design?

At the most basic level, an auditor needs access to data about:

- what the AI system was meant to do
- how it was built to achieve that result
- whether it is, in fact, achieving that result
- the fairness, accuracy, safety and efficacy of the system
- the ongoing monitoring, testing and revision of the system

In order to effectively audit AI systems, developers and deployers of them should maintain and make available documentation for each phase of Ideation, Development and Deployment, such as:

1. Ideation

    a. Purpose of AI system

    b. Expected advantages of AI system over non-AI systems

    c. Potential risks of AI system: severity of those risks and a clear path to managing those risks. Red flags should be identified for each stage in the AI system lifecycle: if a red flag appears then additional work should be paused until the issue can be resolved or, if it cannot be resolved, the AI system should be decommissioned entirely if.

    d. Consultation with groups directly impacted by the use of the AI system. How did you engage those groups, document their concerns, and implement their feedback into the rest of the AI lifecycle?

    e. Plan for human-in-the-loop: Where does a human need to touch the AI system? When does human override of the AI system become necessary?

2. Development

    a. Data used to train and test the model

        i. How collected

        ii. Target selection: what is the AI system optimizing to? How do you quantify that in the AI system?

        iii. Variable selection: what information is passed onto the AI system? How to determine which variables should be included/excluded? Both qualitative and quantitative reasons should be clearly explained.

        iv. Clear identification of training/validation/testing sets of data: how were they selected? How did you prevent data leakage so that the model is not trained with testing data?

    b. Model selection: what motivated the type of model selected? Both qualitative and quantitative reasons should be clearly explained. Model performance metrics and accountability aspects such as transparency should both be considered.

c. Output from AI system: What is the output of the AI system? How is that output used? Does it result in a decision/classification or inform a decision/classification?

d. Industry-standard practices for reproducibility of data and code, including versioning, should be followed and documented.

3. Deployment

a. Plans for continuous testing for validity and accuracy

b. Previously identified performance thresholds that the deployed system is required to meet.

c. Accessibility of the AI system for all users (e.g., taking into account users with different physical abilities, language level, etc.).

d. Transparency: how will the users and affected populations be informed about the decisions made by the AI system?

e. Reporting: If a user/affected party believes an error has occurred, how do they report that?

f. Clear procedures for documenting errors and identifying why the error occurred and how it was remedied.

4. Post-Deployment

a. Set out clear conditions for the AI system to be deprecated and sunsetted.

b. Document the process of removing that AI system from deployment and development pipelines.

c. Identify what, if anything, the AI system will be replaced with and why/how.

Documentation should be retained for as long as the AI system is in development, while it is in deployment, and an additional three years once a system is no longer in active use or development to maintain audit trails and institutional knowledge.

Barriers to Effective Accountability

24. What are the most significant barriers to effective AI accountability in the private sector, including barriers to independent AI audits, whether cooperative or adversarial? What are the best strategies and interventions to overcome these barriers?

Some of the most significant barriers to effective AI accountability in the private sector include:

- Lack of consensus standards - without clear, sector-specific audit frameworks, ad hoc assessments often feel subjective, inconsistent and unfocused. In our experience, many clients are concerned that there isn't a standard they can point to that, if met, provides assurance that their regulatory, litigation and reputational risks are mitigated.

- Protecting competitive advantage - companies may resist oversight that could reveal proprietary methods to competitors. IP concerns.

- Avoiding reputational risks - firms may fear audits surfacing issues that damage their brand image publicly, even if findings are important to address.

- Costs and operational impacts - accountability processes require investments of time, talent and compute resources that impact budgets and schedules.

- Insufficient expertise - many firms lack the skilled, multidisciplinary teams truly needed to conduct meaningful internal or external reviews of complex AI systems.

- Uncertain benefits - hard to measure return on investment may dilute focus on accountability versus more urgent product development and revenue priorities for businesses.

- Compliance fatigue - existing legal obligations make additional requirements feel burdensome unless incentivized and streamlined well.

- Immature tools and methods - issues like AI interpretability, benchmarking, and error analysis remain challenging despite progress, further complicating audits.

Strategies for overcoming barriers and accelerating progress include:

- Voluntary incentives for responsible disclosures, like "bug bounty" programs and safe harbor around findings.

- Industry collaboration to develop common standards and open-source audit tools to reduce costs and maximize learning.

- Integrating accountability into development workflows to minimize disruption, maximize effectiveness.

- Regulatory flexibility to enable customization and focus on principles over prescriptiveness where appropriate.

- User and civil society input mechanisms to guide oversight in addressing broader societal needs beyond company reputation.

- Shared auditing infrastructure like confidential data repositories and compute platforms to defray costs for participants.

- Measuring ROI in terms of risk reduction, trust building, and innovation acceleration alongside costs.

25. Is the lack of a general federal data protection or privacy law a barrier to effective AI accountability?

The lack of a comprehensive federal data privacy and protection law in the US does pose certain barriers to effective AI accountability:

- Fragmented compliance landscape: The patchwork of state, sectoral and common law privacy protections creates confusion and high costs for demonstrating accountability to various requirements. A uniform standard would streamline.

- Uncertainty disincentivizes openness: Fears around litigation risks and lack of clear safe harbors may deter transparency and disclosing issues uncovered through audits. Clearer rules could facilitate more openness.

- Harder to implement strong privacy controls: Accountability for data practices depends on the ability to put rigorous controls in place around access, retention, anonymization, etc. Universal legislation would strengthen AI developers' capabilities on this front.

- Public distrust and scrutiny: The absence of a broad privacy law fuels consumer fears and skepticism about how AI systems utilize their data, necessitating heavy investments in earned trust that clear legislation could help provide.

- No single set of expectations: Varying standards complicate developing consistent audit frameworks and criteria around responsible data practices for AI systems. Aligned federal standards would help.

- FOIA exposure: Lack of standard data privacy exemptions in public record laws may make organizations wary of transparent accountability practices if they could enable access to sensitive data. A federal law could add exemptions.

- Weaker incentives for caution: Without consistent regulatory incentives for data minimization, impact assessments, de-identification requirements, etc., it is easier for firms to take risks with data usage and retention in AI systems. Enforceable rules spur care.

- If properly implemented, such a unified standard could reduce regulatory inefficiency and foster innovation. But an overly burdensome approach could stifle innovation and disadvantage the US compared to other nations.

28. What do AI audits and assessments cost? Which entities should be expected to bear these costs? What are the possible consequences of AI accountability requirements that might impose significant costs on regulated entities? Are there ways to reduce these costs? What are the best ways to consider costs in relation to benefits?

The first level of AI assessment should always be internal, and the creators of the AI system being assessed should bear these costs. These should be seen as simply the cost of doing business in the AI space. An AI system presumably would not make it to deployment if it hadn't proven useful for business purposes. That said, assessing for usefulness in business is a very different task than assessing for fairness, transparency, robustness, and so on.

A well-governed creator of AI systems should have internal assessments in place which do assess for accountability in areas other than business purpose. At the very least, having these measures in place and thoroughly documented will help protect from liabilities such as class actions. Again, the cost is placed on the AI creator.

AI Accountability Policies

31. What specific activities should government fund to advance a strong AI accountability ecosystem?

Continued funding of research in both development and accountability of AI systems is vital. This includes funding for computing power for research universities so that they can compete with the major corporate developers of AI systems.

33. How can government work with the private sector to incentivize the best documentation practices?

- Encourage transparency into the AI system with tools such as model cards
- Provide (via NIST) sector-specific standards and checklists to supplement the NIST AI RMF